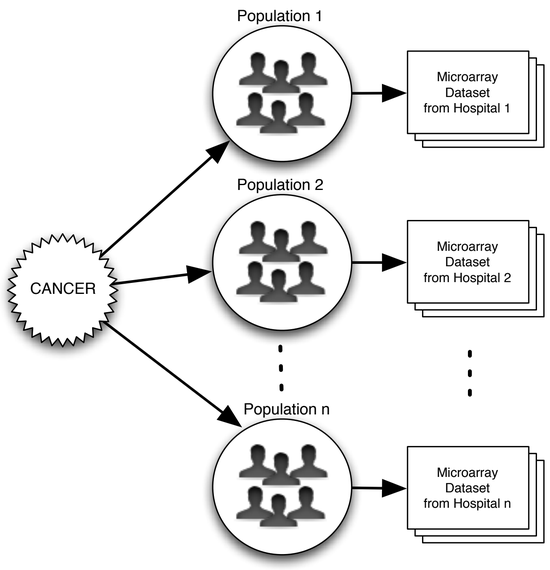
Meta-Analysis in Transcriptional Network Inference

Patrick E. Meyer, Benjamin Haibe-Kains and Gianluca Bontempi

Machine Learning Group
Université Libre de Bruxelles
Belgium

December 2009

Introduction



Problem Given a set of six real breast cancer microarray datasets **(including more than 1000 patients)** coming from different populations, hospitals and microarray platforms (see [1]): **how can we infer a "consensus" transcriptional network** in order to discover new genetic interactions, potentially revealing novel therapeutic targets or prognostic genes?

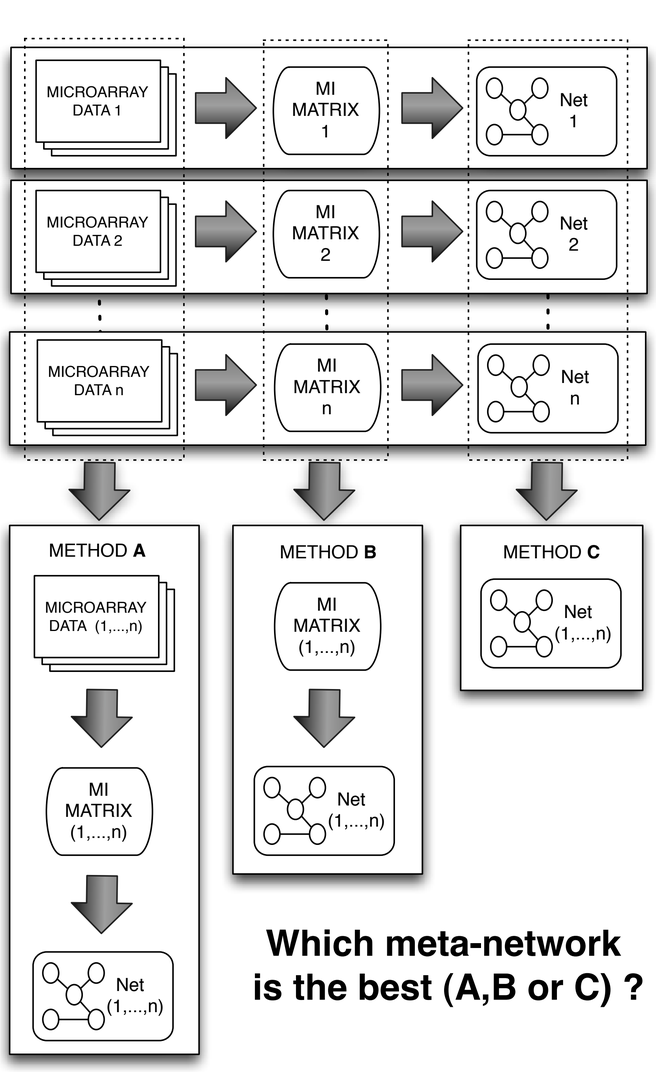
MRNET

MRNET is a transcriptional network inference method particularly adapted for large microarray datasets [2]. MRNET infers a network using the MRMR feature selection method (in a forward selection strategy) where each gene in turn plays the role of the target output X_j.
Given a set X_S of selected variables, the MRMR criterion updates X_S by choosing the variable

$$X_i^{MRMR} = \arg\max_{X_i \in X_{-i}} (u_i - r_i)$$

where $u_i = I(X_i; X_j)$ is a relevance term and $r_i = \frac{1}{|S|} \sum_{X_k \in X_S} I(X_i; X_k)$ is a redundancy term. Hence, MRNET can infer a network from the matrix of pairwise mutual information (MI matrix).
This fast inference method is freely available in the open-source R and Bioconductor package MINET [3].

Meta-network



Which meta-network is the best (A,B or C) ?

Methods

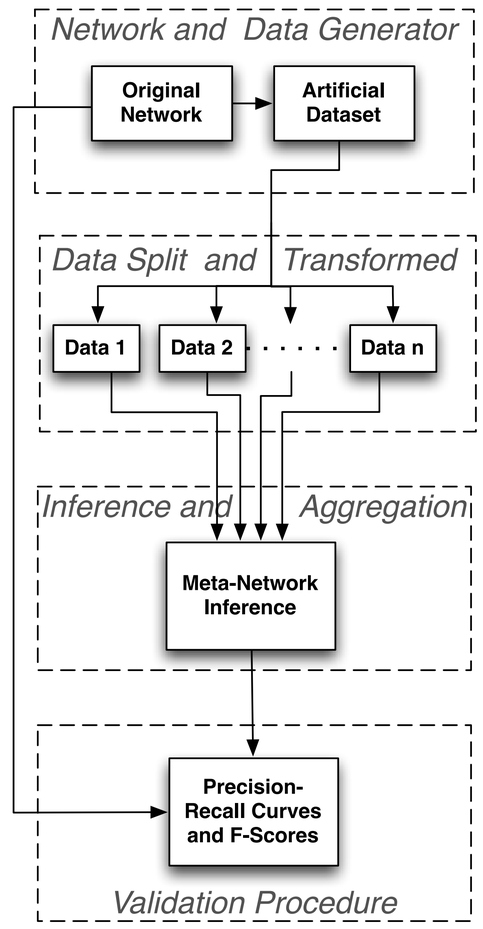
- Method A: Aggregate datasets using standard normalization: $\frac{X-\mu}{\sigma}$
- Method B: Aggregate matrices of pairwise mutual information using a weighted average based on m_j the number of samples in dataset j:

$$\hat{I}^{emp} = \sum_j^n \frac{m_j}{m} \hat{I}_j^{emp}$$

- Method C: Aggregate networks using sum of weights of each arc in each network:

$$W = W_1 + W_2 + ... + W_n$$

Synthetic Experiments

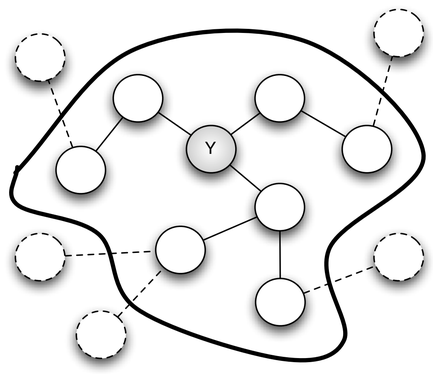


- Artificial Dataset: 300 genes, 800 samples generated with Syntren [2].
- Low heterogeneity configuration:
  - $n = 4$ datasets
  - normally distributed noise
  - random noise intensity between 5% to 15%
- High heterogeneity configuration:
  - $n = 10$ datasets
  - randomly chosen {gaussian, lognormal and gamma} distributed noise
  - with noise intensity between 5% and 30%
  - non-linear transformation of data randomly chosen between {*none*, x^2, $\log(x)$, x^3}

⇒ on 100 runs, method B significantly outperforms (in terms of average F-score) A and C on both configurations.

Real data

Let Y be a multiclass survival variable (indicating time before metastasis), does the set of connected nodes to Y form a predictive signature competitive with previously published ones?



- Six breast cancer datasets
- MRNET applied on each of them
- Meta-network built using Method B
- 100 selected genes connected to survival variable (up to two levels)

Using protocols, signatures and data from [1]

- **The performance of the new signature in a Dataset-CV setting is competitive with the best published prognostic signatures studied.**
- **The selected nodes are highly present in published prognostic signatures representing many different biological processes:**

function	genes	signatures
Proliferation	38	AURKA and GGI
Immune response	4	STAT1 and IRMODULE
Stroma	4	SDPP
Commercial progn.	8	GENE70-76 and ONCOTYPE

References

[1] C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, M. Delorenzi, M. Piccart, and C. Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research*, 14, 2008.

[2] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, Special Issue on Information-Theoretic Methods for Bioinformatics, 2007.

[3] P. E. Meyer, F. Lafitte, and G. Bontempi. Minet: An open source r/bioconductor package for mutual information based network inference. *BMC Bioinformatics*, 2008.